10

Interim Report

ORSER-SSEL Technical Report 6-75

APPLICATIONS OF CLUSTER ANALYSIS IN NATURAL RESOURCES RESEARCH

B. J. Turner

ERTS Investigation 082
Contract Number NAS 5-23133

INTERDISCIPLINARY APPLICATION AND INTERPRETATION OF ERTS DATA WITHIN THE SUSQUEHANNA RIVER BASIN

Resource Inventory, Land Use, and Pollution

Office for Remote Sensing of Earth Resources (ORSER)
Space Science and Engineering Laboratory (SSEL)
Room 219 Electrical Engineering West
The Pennsylvania State University
University Park, Pennsylvania 16802

Principal Investigators:

Dr. George J. McMurtry
Dr. Gary W. Petersen

1082A

Date: July 1975

Made in the United States of America
Reprinted from Forest Science
Vol. 20, No. 4, December 1974
pp. 343–349

Applications of Cluster Analysis in Natural Resources Research

BRIAN J. TURNER

Abstract. Cluster analysis is useful for subsetting a multi-variate multi-observational set of data into homogeneous groups. The technique has wide application as shown by three examples: (a) constructing an abbreviated telephone interview from analysis of responses to a mailed questionnaire; (b) combining tree volume tables over species; and (c) forming spectral signatures from multispectral scanner data. The user is required to exercise judgment in the choice of algorithm, criterion, and number of groups. The selection of the best grouping is therefore subjective. Its greatest value is perhaps as a precursor to more objective analytical techniques. ***Forest Sci. 20:343–349.***

Additional key words. Questionnaire analysis, tree volume, remote sensing.

THE URGE TO CLASSIFY is thoroughly human, and classification is perhaps the most commonly used method of dealing with the unknown. Our first approach in identifying an unfamiliar set of objects is to try to associate it with a known class of objects, thereby reducing the dimensionality of the unknown. If this fails because of the great variation among the unfamiliar objects, it may be instructive to subset these so that objects within a subset are more closely allied with each other than with any member of any other subset. If the subsets cannot still be associated with any known classes, at least it is helpful to know the structure associated with the set so that initial attention can be concentrated on the more important problem of identifying the group, rather than individual differences.

Cluster analysis is concerned with this subsetting problem. The basic premise of cluster analysis is that objects should be placed in the same group if measurements of variables associated with these objects are highly similar. This implies that there should be small variances within a group and large variances between groups. Furthermore if one considered the value of each object as a point in multi-dimensional space, where each dimension represents a measured variable, then objects within a group should be close together and clearly distant from objects in other groups. Alternatively, as in the first application below, it may be desired to subset the variables measured rather than the objects. In either case, the aim is to form subsets of the data that have high internal consistency and maximum separability from other subsets.

Cluster analysis techniques are increasingly being used in taxonomic, ecological, and marketing research. The three applications given here are not drawn from these areas, but represent unusual and very different applications of the technique to classificatory problems in natural resources research.

Numerous algorithms have been suggested for optimally partitioning the data set created by taking one or more measurements on each of a set of objects. Many of these algorithms have been implemented in computer programs and the potential

The author is Associate Professor of Forest Management, School of Forest Resources, The Pennsylvania State Univ. Journal paper no. 4526 of The Pennsylvania State Univ. Agr. Sta., University Park, Pa. Financial support for this research was provided by Hatch funds through research projects 1776 and 1931 of the Pa. Agr. Exp. Sta., State funds through the Dept. of Environmental Resources, and NASA funds through the Office for Remote Sensing of Earth Resources. The Pennsylvania State Univ. Manuscript received Sept. 26, 1973.

user of a cluster analysis technique will probably find that he has to make a choice between programs and then between options within a program. This paper may help the investigator in making these choices because each of the applications described uses a different algorithm. However, I do not suggest that the techniques applied here are the best or only appropriate methods; just that they have given useful results in these cases.

First Application: Questionnaire Analysis

We mailed questionnaires in 1971 to almost 200 land-owners in Warren County, Pennsylvania, in a study of the characteristic and management objectives of land-owners who had recently acquired land in this region (Turner *et al.* 1973).

Only 51 questionnaires were completed and returned. This rather low success rate made it necessary to sample the non-respondents. We did this by telephone, using an abbreviated questionnaire as a basis for interviewing.

In considering the design of the telephone interview, we reasoned that if we performed a cluster analysis on the questions as answered by the mail respondents, we should find certain question-responses that tended to group together. If this were so, then we could presumably infer the answer to any question in the group from the answer to one question within the group.

Educationists involved in testing procedures use the technique of "item analysis" for the similar problem of forming groups of questions ("item pools") so that a test made up of one question from each group will have minimum redundancy in subject matter coverage. We therefore used a computer program prepared for this type of analysis. The program had the additional advantage of simplicity from the user's point of view, making it appropriate for our initial investigations of the technique. The clustering algorithm used is based upon the work of Loevinger *et al.* (1953) who suggested that groups could be built up by first selecting that triplet of variables ("items") that had highest covariances among themselves, then adding variables sequentially to maximize the ratio of the sum of these covariances to the sum of the respective variances. Items that would reduce this "covariance ratio" are discarded. When all items have been considered and either included in the group or discarded, a new group is begun in the same way using the items discarded from the previous group, and so on until all variables have been placed in a group, or a "residue" remains. This algorithm is implemented in a computer program called TEST07 available from the University of Alberta, Edmonton, Alberta, Canada.

By considering the coded responses to the questionnaire questions as being the variables in this method, we were able to obtain groupings of questions-responses. After selecting at least one question from each group, we were able to formulate a much-abbreviated questionnaire suitable for telephone interviews. From the reponses to these questions and considering the groupings of the questions, we felt confident that we could infer the answers to the omitted questions.

Second Application: Volume Table Construction

In preparing a set of volume tables for the commercial forest species of Pennsylvania (Turner 1972a), I developed cubic-foot and board-foot volume equations for 21 species or species-groups. Where these equations are used in computer programs for volume computations, there is no real advantage in amalgamating species equations; and there would undoubtedly be some loss of precision. However, for less precise field use there could be some advantage in reducing the number of tables to a smaller set.

The models for estimating volume were of the form:

$$V = b_1 + b_2 D^2 H$$

for board-foot volume, and

$$V = b_1 + b_2 D^2 H$$

for cubic-foot volume,

where

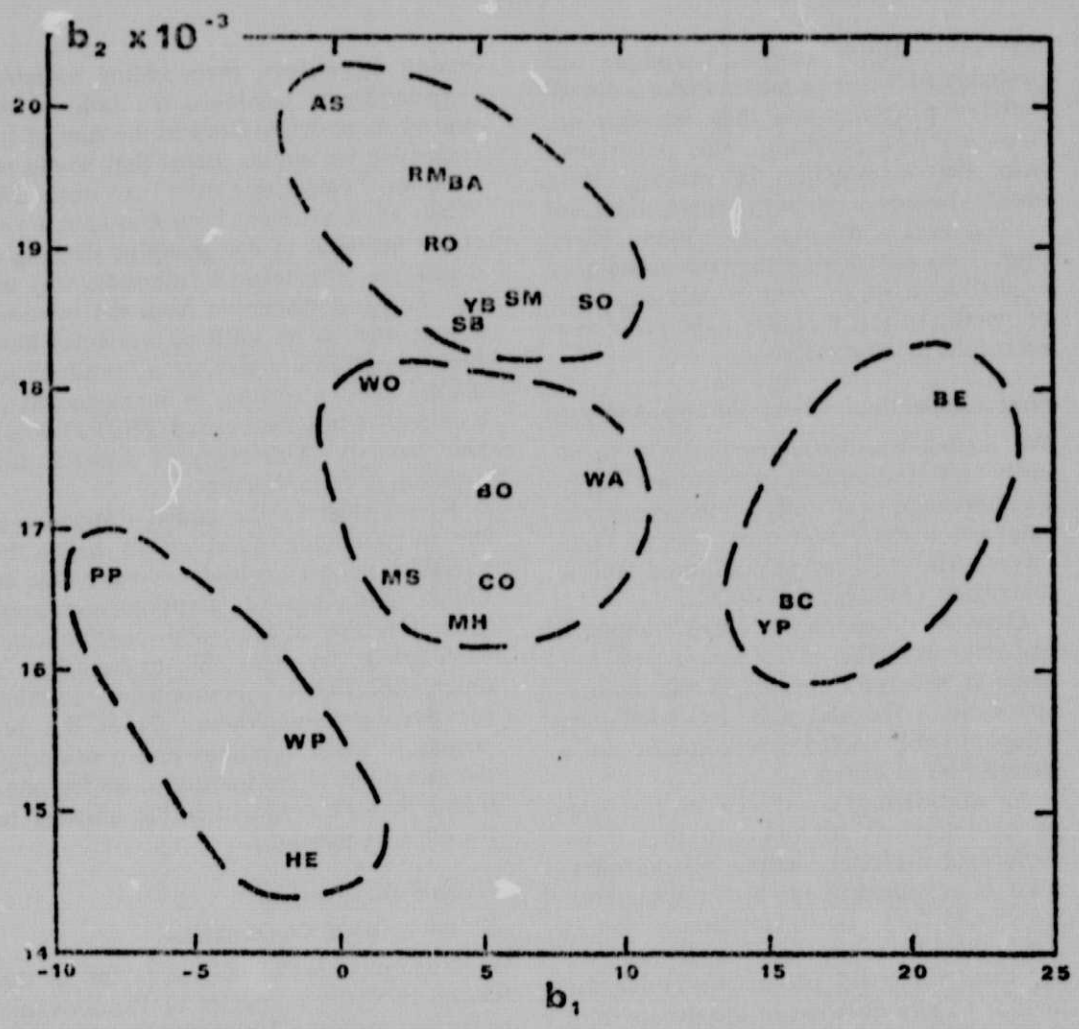

FIGURE 1. *Scatter diagram of board-foot volume coefficients (b_1 and b_3) for different species, and the four groups as defined by cluster analysis (see Table 1 for interpretation of species symbols).*

$\hat{V}$ = predicted volume,
D = dbhob,
H = the appropriate merchantable height, and the
b_i = estimated regression coefficients.

A suitable basis for grouping species equations would therefore be the estimation parameters, b_i. The cluster analysis program described by Rubin and Friedman (1967) was used. This program is actually a comprehensive package of clustering algorithms with options for considering continuous or discrete data and for using a number of different criteria.

This method of subsetting continuous data (also described by Friedman and Rubin 1967) is based on the partitioning of the total scatter matrix,

$T = X'X$ where X is the $n \times p$ matrix of n observations on p variables standardized to mean zero and variance = 1,

into g groups such that some scalar property of the pooled within-group scatter matrix (W) or of the between-groups scatter matrix (B) is optimized. The choice of the number of groups, g, and the

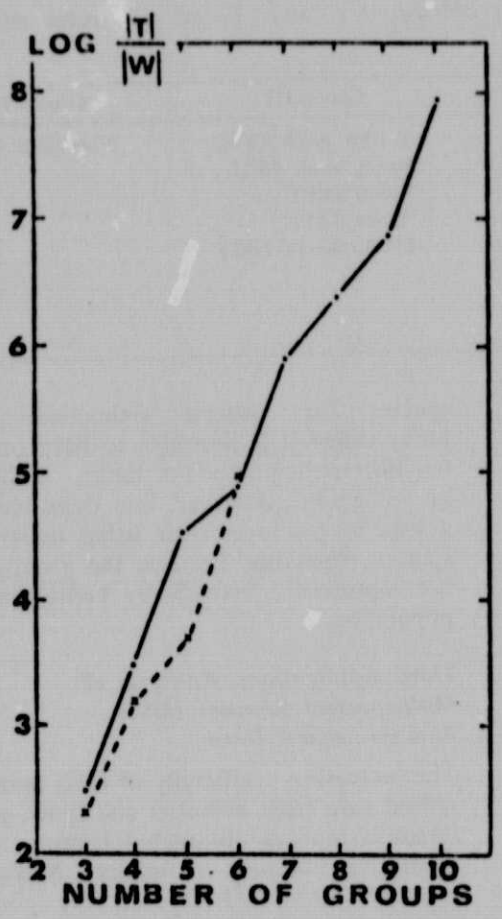

FIGURE 2. *Logarithm of cluster analysis criterion, (|T|/|W|), plotted against number of groups formed for both board-foot volume coefficients (solid line) and first two principal components of cubic-foot volume coefficients (dashed line).*

most appropriate criterion is left to the user.

This poses one of the dilemmas of cluster analysis techniques. Occasionally the user may know how many groups into which he wants to subset the data; more often he may have a desirable range of g-values. The clustering methods will not give the best value of g, although sometimes this may be indicated by a sudden increase in the optimal criterion value as g is varied. Furthermore, different criteria for a given g-value give differing group compositions. The user may wish to try different criteria and look for consistent groupings, and temper this with his intuitive feel for the structure of the data before making final groupings. Alternatively, he may rely on the findings of Scott and Symons (1971) and Marriott (1971) that for many types of continuous data the best criterion is to minimize the determinant of W ($|W|$).

In the case of the volume table analysis, I decided that fewer than four species-group tables would not cover the variation adequately, whereas more than six would negate the value of a reduced set of tables. Several different criteria were used; only some are reported here.

The board-foot volume coefficients were analyzed first because the fact that only two variables are involved permitted a ready comparison between the solutions and a scatter plot of b_2 against b_1 (Fig. 1). The criterion used was: maximize ($|T|/|W|$), which is equivalent to minimizing $|W|$ because $|T|$ is constant for a data set. When the logarithms of the maximum criterion values for different g values were plotted against g, it was evident that the increases in the criterion value from $g = 3$ to 4 and $g = 4$ to 5 were similar and substantially greater than the increase from $g = 5$ to 6 (Fig. 2). This gave an objective basis for rejecting $g = 6$. Assuming that, all other things being equal, the least number of volume tables, the better; $g = 4$ was preferred over $g = 5$. The optimum composition of the groups for $g = 4$ is shown in Figure 1.

Because there were significant intercorrelations among the four cubic-foot volume coefficients, a reduction in computer processing time seemed possible by first taking principal components (an option available with this program) and then clustering a reduced set of components. The first two components explained 89 percent of the variation; hence cluster analysis was performed on these. By using the ($|T|/|W|$) criterion again, maximum values were obtained for varying g-values and their logarithms plotted against g (Fig. 2). This indicated that $g = 4$ or 6 were suitable values, so again, four groups were chosen. The composition of the groups is given in Table 1.

TABLE 1. Optimal species grouping for cubic-foot volumes based on cluster analysis of equation coefficients.

Group I	Group II	Group III	Group IV
White pine (WP)	Red pine (RP)	Yellow birch (YB)	Pitch pine (PP)
Hemlock (HE)	Sugar maple (SM)	Sweet birch (SB)	
Misc. softwoods (MS)	Red maple (RM)	Beech (BE)	
Black oak (BO)	Red oak (RO)	Aspen (AS)	
Chestnut oak (CO)	Scarlet oak (SO)	Black cherry (BC)	
Yellow poplar (YP)	White oak (WO)		
Misc. hardwoods (MH)	White ash (WA)		
	Basswood (BA)		

As a check on the groupings, cubic-foot volumes were read from the tables for three representative dbh's and heights. Mean volumes and standard deviations were computed for each of the three tree sizes of each group (Table 2). Although the standard deviations are somewhat overlapping, there does appear to be on this basis more difference between groups than within them.

The normal statistical procedure for testing whether simple linear regression models might be combined is to test for slope and intercept differences in an analysis of variance (for example, Williams 1959), and combine models on the basis of multiple range tests on the regression coefficients. In the case of the board-foot volume models, the differences between slope coefficients (b_2) were so large in comparison with their standard deviations that such a procedure would have reduced the number of equations only by about two. If a similar statistical procedure were available for testing the non-linear cubic-foot volume models, a similar situation would likely be found. Cluster analysis provides a means, therefore, for grouping species for volume estimation purposes where it is desirable to have only a few different equations or tables. It should be recognized, however, that there will be a loss in precision over using individual species equations, because the groups do not represent a statistically homogeneous population.

Third Application: Analysis of Multispectral Scanner (MSS) Remote-Sensed Data

The increasing availability of MSS remote-sensed data from airborne and space platforms is making automated land-use and vegetative mapping a reality. Mapping methods require that representative spectral signatures be obtained for each target of interest (land-use class, species type, and so on) and then each MSS response element be classified according to which target signature is nearest.

A response element on a digital tape derived from MSS data consists of the spectral reponse (or spectral signature) from a point on the ground. The response is a vector of size equal to the number of bands into which the visible and near-visible

TABLE 2. Cubic volumes for three tree sizes, with 21 species equations clustered into four groups (converted to metric units).

Dbh (cm)	Merchantable height (in 2.4 m bolts)	Mean ± Standard Deviation (m^3)			
		Group I (7 species)	Group II (8 species)	Group III (5 species)	Group IV (1 species)
12.7	2	0.053 ± 0.006	0.056 ± 0.006	0.062 ± 0.006	0.022
35.6	7	.815 ± .022	.862 ± .020	.907 ± .031	.823
61.0	10	2.965 ± .076	3.254 ± .120	3.632 ± .238	3.349

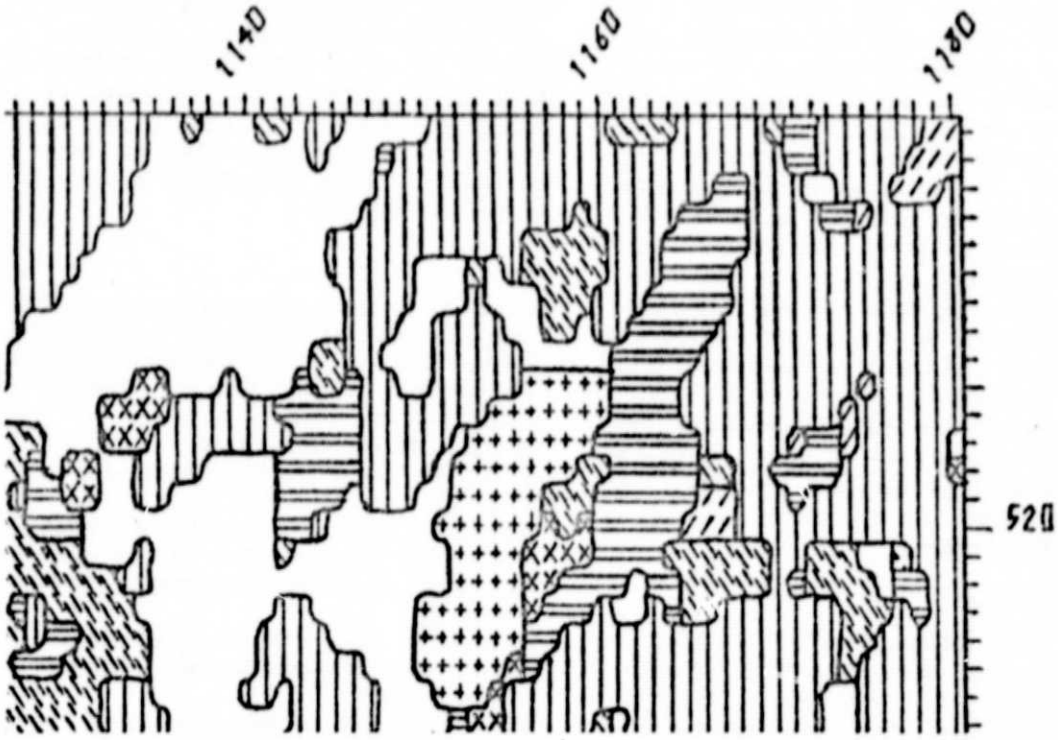

FIGURE 3. *Portion of plotter line map of Stone Valley Experimental Forest and surroundings, Huntingdon County, Pennsylvania, derived from computer processing of ERTS-1 satellite MSS data. The area shown represents about 600 ha. Coordinates refer to scan line and element numbers. Types depicted are: hardwoods (vertical lines), shaded hardwoods (right-to-left slashes), hemlock-hardwood (horizontal lines), conifers (left-to-right slashes), water (+'s), fields (open areas), unclassified (X's).*

spectrum is split by the scanner. The value of the vector is the intensity of radiation received by the scanner in each spectral band from the geographical point. The size of the point varies with the altitude of the scanner and other factors; for ERTS–1 satellite data it is about 0.5 ha. The data can be arranged on a digital tape in geographic relationship so that a scene of interest can be selected by referencing the boundary scan lines (perpendicular to the flight line) and elements within scan lines.

One method for obtaining representative spectral signatures is cluster analysis. As a component of a system of computer programs for analyzing MSS digital tapes (Borden 1972), I wrote a program using a cluster analysis algorithm (Turner 1972b) which is now being used in the analysis of data from ERTS–1 as well as aircraft data.

The clustering algorithm was influenced by the "iterative condensation on centroids" procedure (Tryon and Bailey 1970), a useful method when the number of observations is very large. The method has, however, been freely adapted for computational efficiency, since we need only the mean and variance of each group or target-type and not the elemental composition of each group.

The algorithm uses the first scan line of the scene of interest to form trial centroids or mean spectral signatures. The number of these formed is controlled by the user who specifies a critical separation value for the formation of a new centroid. The whole scene is then sampled and the mean spectral signatures are revised, and added to if necessary, on the basis of the sampled elements and variances computed from the data. Thus while the intensity of

classification is under user-control, variability in the sample data is used to assist in the assigning of elements to groups. For instance, a "mixed hardwoods" group will have more inherent variability than a "pure conifer" group. The output from the program is thus a set of mean spectral signatures with a measure of the variability within each group. The scene can then be reprocessed to produce a line printer character map with characters assigned according to the computed mean spectral responses and critical values derived from the variability measures. Alternatively, the output data can be combined with spectral data derived by other means, and character or plotter-drawn line maps can be produced. An example of the latter is shown in Figure 3.

Conclusions

The three foregoing applications illustrate the potential of cluster analysis techniques in dealing with a wide range of classification problems. They also illustrate some of the difficulties associated with using cluster analysis:

(a) There are a large number of different methods associated with the general term "cluster analysis" and the user must exercise care in choosing an appropriate one.
(b) The user is generally required to make some judgmental decisions in using the technique, for instance, in the choice of number of groups desired.
(c) The user will generally have to experimentally vary parameters and finally have to make a subjective judgment as to which result is "best."

Despite this, cluster analysis provides a relatively easy and cheap method of taking a first look at multivariate data. Because of the wide variety of clustering algorithms and their general robustness, continuous, discrete, or mixed data can be analyzed with equal facility. The detection of groupings within the data set may provide suggestions for hypotheses which can then be tested by more refined and objective analytical techniques.

Literature Cited

BORDEN, F. Y. 1972. A digital processing and analysis system for multispectral scanner and similar data. *In*: Remote Sensing of Earth Resources 1:481–500. Univ of Tennessee Space Institute. 783 p.

FRIEDMAN, H. P., and J. RUBIN. 1967. On some invariant criteria for grouping data. J Am Stat Assoc 62:1159–1178.

LOEVINGER, J., G. C. GLESER, and P. H. DUBOIS. 1953. Maximizing the discriminating power of a multiple-score test. Psychometrika 18(4): 309–317.

MARRIOTT, F. H. C. 1971. Practical problems in a method of cluster analysis. Biometrics 27(3):501–514.

RUBIN, J., and H. P. FRIEDMAN. 1967. A cluster analysis and taxonomy system for grouping and classifying data. IBM Contributed Program Library. 221 p.

SCOTT, A. J., and M. J. SYMONS. 1971. Clustering methods based on likelihood ratio criteria. Biometrics 27(2):387–397.

TRYON, R. C., and D. E. BAILEY. 1970. Cluster analysis. McGraw-Hill, New York. 347 p.

TURNER, B. J. 1972a. Board-foot and cubic-foot volume tables for the commercial forest species of Pennsylvania. Coll of Agric, The Pennsylvania State Univ. 69 p.

———. 1972b. Cluster analysis of multispectral scanner remote sensor data. *In*: Remote Sensing of Earth Resources 1:538–549. Univ of Tennessee Space Institute. 783 p.

———, C. H. STRAUSS, and L. J. SWANDIC. 1973. Pennsylvania land buyers are young, affluent, and well educated. J Forest 71(12): 770–772.

WILLIAMS, E. J. 1959. Regression analysis. Wiley. 214 p.